Automated data labeling of building automation systems using time series data and conditional probabilities

M Maghnie[1], F Stinner[1], A Kümpel[1] and D Müller[1]

[1] RWTH Aachen University, E.ON Energy Research Center, Institute for Energy Efficient Buildings and Indoor Climate, Mathieustraße 10, 52072 Aachen, Germany

E-mail: marwa.maghnie@eonerc.rwth-aachen.de

Abstract. As energy efficiency demands increase, buildings get smarter, and the amount of data to analyze grows, where each building device may generate multiple data streams. These extensive quantities of monitoring data serve as a great opportunity for detecting anomalies in building automation systems and for optimizing their control. However, each building usually uses a custom format for data labels, therefore requiring an individual data label analysis per building. This makes the conceptually manageable task of detecting energy systems from the raw data increasingly complex and error-prone, which is a further hurdle that any building operation optimizer must resolve. This paper presents a methodology for automatically categorizing and labeling raw monitoring data from building automation systems. Using statistical features of the data, the method checks which data streams follow which known building operation rules and patterns. Therefore, an initial labeling of the data streams takes place. Furthermore, examining the correlation between the data streams indicates possible related system components using the concept of conditional probability. As a use case for the methodology, unlabeled data from a real building automation system are examined. The results show that, using unlabeled time series, data types from certain sensors and actuators can be reliably identified. The proposed methodology could therefore simplify the implementation of energy applications such as operation optimization and fault detection of building automation systems

1. Introduction

The operation of building energy systems in the age of Industry 4.0 leads to immense amounts of data that usually have to be evaluated before they can be used to improve energy efficiency and analysis. However, against the advantages of digitalization come the challenges of having to sort out these data streams that often have different naming conventions across different buildings, often without an intuitive interpretation of the data label meanings. Therefore, data-based energy analysis is generally limited by challenging data sets [2, 5]. Especially unlabeled data sets, or those with cryptic labels, pose a challenge when it comes to useful applications, such as fault detection and diagnostics [6]. In order to use the data to increase energy efficiency of building control, experts usually have to identify the data manually, i.e. they have to sort through building plans and make numerous visitations to locate which building component is producing which data stream. The use of time series classification (TSC) algorithms can help alleviate these challenges. TSC algorithms can be divided into three main categories: deep learning, statistical and distance-based [1]. Available deep learning methods usually require big

amounts of data (e.g. 10,000 data streams or more) [7, 8]. Furthermore, these deep learning methods are usually not applicable on time series data from new or different buildings, or require extensive time and effort in training models. Especially considering smaller data sets (e.g. 1000 data streams or less), the benefits of implementing deep learning methods may not outweigh the efforts. Therefore, to prevent the extra effort for the user and to make the process more automatic, we focus on statistical and distance-based heuristics in this work. Furthermore, only unlabeled time series data are required as input to the proposed methodology in this work. Although it was shown in related work that using data stream labels significantly increases chances of correct classifications, the methodology would not likely be applicable to all building operators [8]. Furthermore, most existing methodologies that use data labels as well do not focus on examining the inter-component relationships between the time series data [11]. However, this work considers the correlations between classified data streams to detect related subsystems as well.

1.1. Related work

McArthur and El mokhtari propose a hybrid framework that uses both rule-based and AI-based classification [1]. Their method also estimates certain missing time series, based on the known time series values. However, they do not consider or examine the effect of conditional probabilities in their work. Mertens and Wilde focus only on using time series data as well and reach advanced results. They use error probabilities but do not consider yet conditional probabilities to improve the classification [9]. Ma et al. used an approach similar to conditional probability to increase the accuracy of label predictions, but focused on classifying only sensor data [10]. Balaji et al. also experimented with using only time series data and some inter-component relationships, yet for their main focus incorporated data labels [11]. As the majority of current work in TSC of building energy system data uses semantic interpretation of data point labels, the present work could be vital in scenarios where data point labels are not interpretable.

2. Methodology

As outlined in Figure 1, the methodology requires only time series data as input. Before the TSC phase, the time series data is possibly reformatted and cleaned in a preprocessing step. Following the data reformatting, certain statistical functions for all the data are calculated. Using statistical functions is almost an ubiquitous element in TSC algorithms to detect related components, such as the work of Hong et al. and Stinner et al. [12, 4]. The specific selection of statistical functions can differ according to the specific use cases. Statistical values allow recognizing uncategorizable data, such as data points that have a constant value (e.g. of 0 or 1). These data points do not provide much information to allow extracting meaningful features and they would only add unnecessary processing load on the methodology. Therefore, these data points are filtered out. Furthermore, using the statistical summary data, we can check

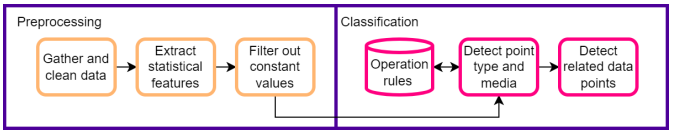


Figure 1: Methodology flow includes a preprocessing step for data preparation that is followed by the actual classification of the time series

for certain known (operation) patterns. These operation rules and limits for expected values are taken from aggregations of data sets for building automation systems during normal, i.e. non-faulty, operation, such as those by Stinner et al. [3]. For example, control devices usually

send their commands as a percentage value, to indicate e.g. how much a valve should open. Therefore, given that x is a single time series, then fulfilling the conditions $min(x) \approx 0$ and $max(x) \approx 100$ is a strong indicator that x is a control value. The methodology optimistically assumes that once a pattern is fulfilled by a time series, the time series is 100 % likely to be that particular assignment. Another example of these heuristically created rules would be the composite condition $(-10 \leq x \leq 40)|(283 \leq x \leq 363)$ which would be a sign that x represents a temperature (air or water), considering both the degree Celsius limits and the Kelvin limits.

These were examples of independent conditions or rules. However, more insights into the data can be revealed when considering conditional rules, where a classification depends on the relationship between two or more (possibly already classified) time series. For example, the rule $\texttt{correlation}(sensor_x, setpoint_y) \geq 0.97$ indicates that $sensor_x$ is directly affected by $setpoint_y$, given that these data points have been independently recognized first as a sensor measurement or a set point. However, in this work, this is only considered as an indication and not as a definitive proof of an existing relationship between the data points. To allow the configuration and application of the operation rules, these rules are implemented and subsequently evaluated using SymPy, the open source symbolic mathematics library [13].

3. Experimental setup

To test the method, it is applied on a small data set containing 18 data streams from multiple HVAC systems in a real university building. For some variety of data types, the data streams include those from heater valves, cooler valves, and thermal zone air temperatures. As for the data resolution, it is variable and ranges from about a reading every minute to every 5 minutes. The data set spans over three months, from January to March, 2021. Table 1 shows a sample of the monitoring data stream labels and their meanings. An understandable naming convention could not be determined.

Table 1: Sample of the actual data stream labels and their meanings ("[...]" is used to replace identical prefixes)

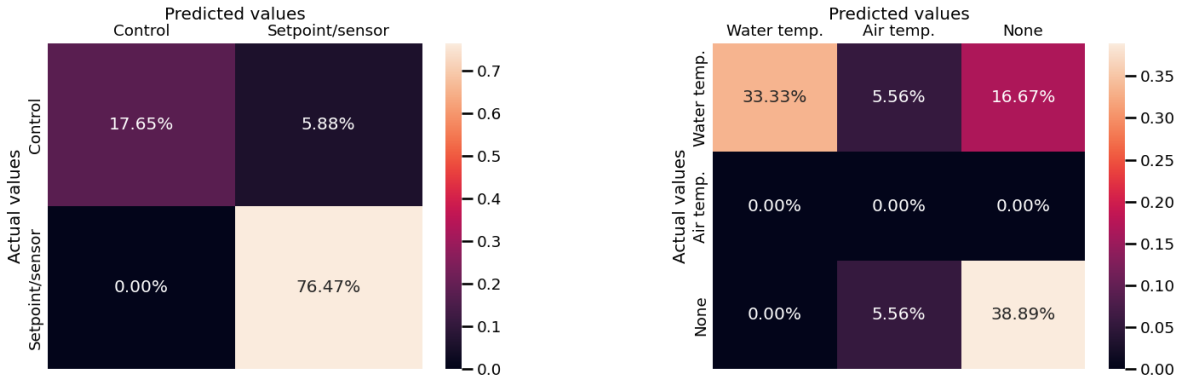
Data stream label	Meaning
`[...]000/01/08/H614.01/6-AV2796220`	Valve control signal of heater 1
`[...]001/00/00/P.08/CAI/b-AI2796210`	Valve position of heater 1
`[...]206/00/00/P.16/CAI/b-AI326030`	Outlet temperature of heater 1
`[...]001/00/00/P.05/CAI/b-AI2796209`	Inlet temperature recirculation of heater 1
`[...]000/05/08/H614.01/6-AV2796286`	Valve control signal of heater 2
`[...]203/00/00/P.08/CAI/b-AI2796234`	Valve position of heater 2
`[...]206/00/00/P.19/CAI/b-AI326066`	Outlet temperature of heater 2
`[...]203/00/00/P.05/CAI/b-AI2796233`	Inlet temperature recirculation of heater 2

4. Experimental results and discussion

As shown in Table 2, the methodology begins by calculating meaningful statistical functions. These specific features were chosen because they are needed in the heuristically defined operation rules of building automation. Afterward, the classification is executed, and the results are plotted in Figure 2. After applying the operation rules and testing if the patterns are fulfilled, the general categorization classifications are mostly correct. In the first step of point type detection

Table 2: Features of the time-series data

Data point (unit)	Mean	Std. Dev.	Minimum	Maximum
Valve control signal LE01 (%)	22.74	32.97	0.00	100.00
Valve position LE01 (%)	22.53	32.99	0.24	100.63
Outlet temperature LE01 (°C)	28.46	12.57	15.33	87.34
Inlet temperature recirculation LE01 (°C)	24.91	12.28	12.23	84.92
Valve control signal L02 (%)	31.40	42.14	-0.49	100.00
Valve position L02 (%)	-0.16	0.57	-0.72	5.68
Outlet temperature L02 (°C)	36.07	16.43	18.31	91.12
...	...	...	...	...
Valve control signal LK01 (%)	0.00	0.00	0.00	0.00
Valve position LK01 (%)	0.81	2.73	0.17	68.78
Outlet temperature LK01 (°C)	18.64	2.95	10.06	27.46
Inlet temperature recirculation LK01 (°C)	42.47	15.37	17.77	100.95
Valve control signal H04 (%)	99.99	0.41	0.00	100.00
Outlet temperature H04 (°C)	44.05	7.56	20.00	67.74
Inlet temperature recirculation H04 (°C)	42.62	5.37	7.55	59.37



(a) Actual vs. predicted point types

(b) Actual vs. predicted media type

Figure 2: Confusion matrices of the time series classification

(Figure 2a), almost 76 % + 17 % = 93 % of the classifications were true positives. The second step, which comprises media type classification (Figure 2b), continued to show mostly positive results, yet at a reduced capacity of around 73 %. The reduced accuracy is partly due to faulty monitoring data, which does occur in practice. As shown in Table 2, some data points exhibit unexpected features. For example, in the second section, "Valve position L02" includes negative values, and in the fourth section, "Valve control signal LK01" is completely missing from the monitoring platform (the latter, however, was already filtered out before the classification). Since the proposed method relies solely on time series data, faulty data inhibit the capabilities of the method to identify actual system features and detect inter-component relationships. However, quantifying the effects of faulty data on TSC is not within the scope of the current work.

Finally the methodology detects all the piece-wise correlation values between the time series, as high values can indicate data streams that directly affect each other, i.e., they might belong to the same system component. For example, simply as a visual aid for the reader, when the data are plotted, a pattern presents itself in some cases (Figure 3).

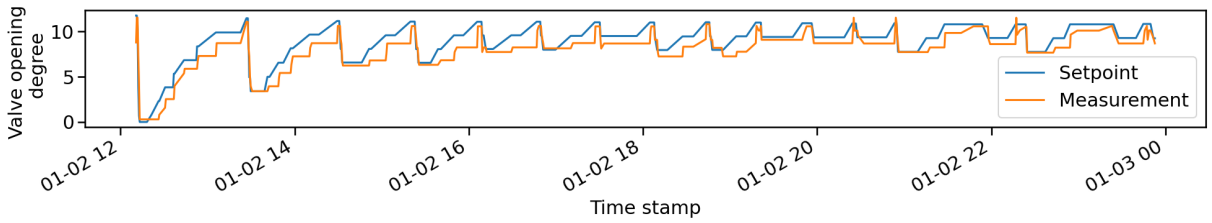
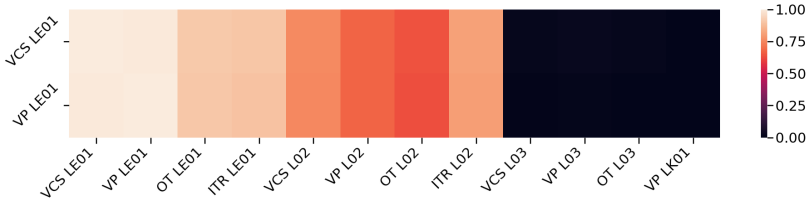
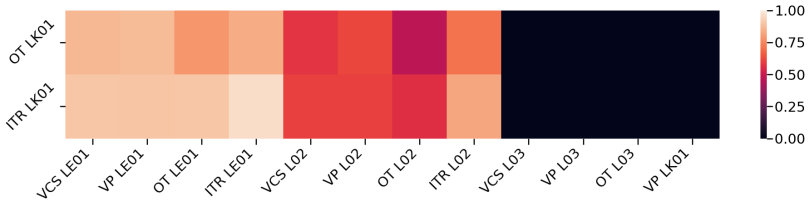


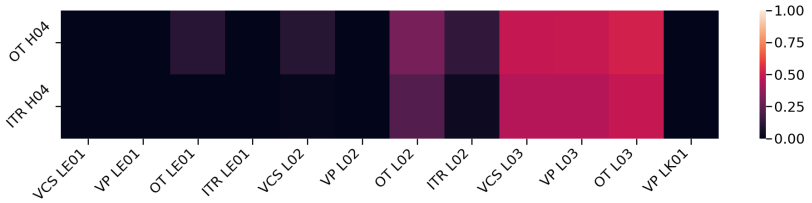
Figure 3: The relationship between the valve opening set point and the measurement values can be used to generate conditional probabilities of classifications



(a) Correlation between data streams in the same component vs. in different components



(b) Correlations with a clear pattern towards zone location



(c) Very low correlations between distant data streams

Figure 4: Correlations between time series (VCS: Valve Control Signal; VP: Valve Position; OT: Outlet Temperature; ITR: Inlet Temperature Return)

Cross-checking the initial independent classifications with the correlations values (Figure 4) allows classifying subsystems while considering conditional probability. For example, in Figure 4a, it is clear that the data stream of the valve control signal from the air heater in zone 1 (VCS LE01) has the highest correlation with its corresponding valve position data stream (VP LE01). With data streams in other subsystems and zones, the correlation decreases significantly. However, an outlier is demonstrated in Figure 4b, where the air cooler data streams (OT LK01 and VP LK01) do not show any correlation. Finally in Figure 4c, the same pattern of correlation vs. zone placement is shown from the perspective of data streams in zone 4 (i.e. OT H04 and ITR H04). These correlation results could be helpful in detecting the related subsystems in the

building. However, further studies need to be carried out to fully investigate the implications of correlated data points.

A current challenge in the methodology is the sensitivity of the operation rules against outliers in the time series values. Possible solution approaches include allowing a certain percentage of a time series to not fit to the operation rules, for example, using multi-class support vector machines. This margin of error approach, and generally a more comprehensive handling of faulty data, is one possible direction for further developing the presented prototype framework.

5. Conclusion and outlook

In this paper we describe a method for classification of building automation data using only the time series data, without the effort and time in model training. Applying heuristic rules of operation from building automation systems, the unlabeled time series from a real building were classified into general categories with an average accuracy of 83 %. In order for the time series classification method introduced in this work to be usable at a wider scale, transfer learning can be a future direction. Hong et al. have demonstrated a prototype in this direction, using both time series data and their labels [12]. Applying the proposed methodology on a larger data set could also help in quantifying its scalability. Furthermore, the list of operation rules can be expanded to consider more data types, such as those related to humidity levels or general air quality. By automating the process of data labeling in buildings, applications such as energy efficiency analysis of building control can be applied in a scalable manner to a large number of buildings, making a significant contribution to reducing CO2 emissions in the future.

Acknowledgment

We gratefully acknowledge the financial support provided by the BMWK (Federal Ministry for Economic Affairs and Climate Action), promotional reference 03EN1014A.

References

[1] McArthur JJ, El mokhtari K. A data-driven approach to automatically label BAS points [Internet]. 2021
[2] Stinner F, Kornas A, Baranski M, Müller D 2018 Structuring building monitoring and automation system data
[3] Stinner F, Kümpel A, Müller D 2023 Automated data analysis for scalable application in existing buildings
[4] Stinner F, Gorgis D, Kümpel A, Müller D 2023 Automatic detection of control loops in existing buildings
[5] Bode G, Schreiber T, Baranski M, Müller D 2019 A time series clustering approach for Building Automation and Control Systems. Applied Energy. Mar;238:1337–45.
[6] Matetić I, Štajduhar I, Wolf I, Ljubic S 2023 A Review of Data-Driven Approaches and Techniques for Fault Detection and Diagnosis in HVAC Systems. Sensors. 2023 Jan;23(1):1.
[7] Fütterer J, Kochanski M, Müller D. Application of selected supervised learning methods for time series classification in Building Automation and Control Systems. Energy Procedia. 2017 Sep 1;122:943–8.
[8] Stinner F, Llopis-Mengual B, Storek T, Kümpel A, Müller D 2022 Comparative study of supervised algorithms for topology detection of sensor networks in building energy systems. Automation in Construction.
[9] Mertens N, Wilde A 2023 Automated Classification of Datapoint Types in Building Automation Systems Using Time Series.
[10] Ma J, Hong D, Wang H 2020 Selective sampling for sensor type classification in buildings. In 2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN) (pp. 241-252)
[11] Balaji B, Verma C, Narayanaswamy B, Agarwal Y 2015 Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for Buildings.
[12] Hong D, Wang H, Ortiz J, Whitehouse K. 2015 The Building Adapter: Towards Quickly Applying Building Analytics at Scale.
[13] Meurer A, Smith CP, Paprocki M, Čertík O, Kirpichev SB, Rocklin M, Kumar A, Ivanov S, Moore JK, Singh S, Rathnayake T. SymPy: symbolic computing in Python. PeerJ Computer Science. 2017 Jan 2;3:e103.